

Statistical Matching using Fractional Imputation

Jae-kwang Kim

Emily Berg

Taesung Park

October 14, 2015

Abstract

Statistical matching is a technique for integrating two or more data sets when information available for matching records for individual participants across data sets is incomplete. Statistical matching can be viewed as a missing data problem where a researcher wants to perform a joint analysis of variables that are never jointly observed. A conditional independence assumption is often used to create imputed data for statistical matching.

We consider an alternative approach to statistical matching without using the conditional independence assumption. We apply parametric fractional imputation of Kim (2011) to create imputed data using an instrumental variable assumption to identify the joint distribution. We also present variance estimators appropriate for the imputation procedure. We explain how the method applies directly to the analysis of data from split questionnaire designs and measurement error models.

Key Words: Data combination, Data fusion, Hot deck imputation, Split questionnaire design, Measurement error model.

1 Introduction

Survey sampling is a scientific tool for making inference about the target population. However, we often do not collect all the necessary information in a single survey, due to time and cost constraints. In this case, we wish to exploit, as much as possible, information already available from different data sources from the same target population. Statistical matching, sometimes called data fusion (Baker et al., 1989) or data combination (Ridder & Moffit, 2007), aims to integrate two or more data sets when information available for matching records for individual participants across data sets is incomplete. D’Orazio et al. (2006) and Leulescu & Agafitei (2013) provide comprehensive overviews of the statistical matching techniques in survey sampling.

Statistical matching can be viewed as a missing data problem where a researcher wants to perform a joint analysis of variables that are never jointly observed. Moriarity & Scheuren (2001) provide a theoretical framework for statistical matching under a multivariate normality assumption. Raessler (2002) develops multiple imputation techniques for statistical matching with pre-specified parameter values for non-identifiable parameters. Lahiri & Larsen (2005) address regression analysis with linked data. Ridder & Moffit (2007) provide a rigorous treatment of the assumptions and approaches for statistical matching in the context of econometrics.

	X	Y_1	Y_2
Sample A	o	o	
Sample B	o		o

Table 1: A Simple data structure for matching

Statistical matching aims to construct fully augmented data files to perform statistically valid joint analyses. To simplify the setup, suppose that two surveys, Survey A and Survey B, contain partial information about the population. Suppose that we observe x and y_1 from the Survey A sample and observe x and y_2 from the Survey B sample. Table 1 illustrates a simple data structure for matching. If the Survey B sample (Sample B) is a subset of the Survey A sample (Sample A),

then we can apply record linkage techniques (Herzog et al. , 2007) to obtain values of y_1 for the survey B sample. However, in many cases, such perfect matching is not possible (for instance, because the samples may contain non-overlapping subsets), and we may rely on a probabilistic way of identifying the “statistical twins” from the other sample. That is, we want to create y_1 for each element in sample B by finding the nearest neighbor from Sample A. Nearest neighbor imputation has been discussed by many authors, including Chen & Shao (2001) and Beaumont & Bocci (2009), in the context of missing survey items.

Finding the nearest neighbor is often based on “how close” they are in terms of x ’s only. Thus, in many cases, statistical matching is based on the assumption that y_1 and y_2 are independent, conditional on x . That is,

$$y_1 \perp y_2 \mid x. \tag{1}$$

Assumption (1) is often referred to as the conditional independence (CI) assumption and is heavily used in practice.

In this paper, we consider an alternative approach that does not rely on the CI assumption. Instead, we adopt an approach to statistical matching based on an instrumental variable, as discussed briefly in Ridder & Moffit (2007). Kim & Shao (2013) propose the fractional imputation method for statistical matching under an instrumental variable assumption. After we discuss the assumptions in Section 2, we review the fractional imputation methods in Section 3. Furthermore, we consider two extensions, one to split questionnaire designs (in Section 4) and the other to measurement error models (in Section 5). Results from two simulation studies are presented in Section 6.

2 Basic Setup

For simplicity of the presentation, we consider the setup of two independent surveys from the same target population consisting of N elements. As discussed in Section 1, suppose that Sample A collects information only on x and y_1 and Sample B collects information only on x and y_2 .

To illustrate the idea, suppose for now that (x, y_1, y_2) are generated from a normal distribution such that

$$\begin{pmatrix} x \\ y_1 \\ y_2 \end{pmatrix} \sim N \left[\begin{pmatrix} \mu_x \\ \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{xx} & \sigma_{1x} & \sigma_{2x} \\ & \sigma_{11} & \sigma_{12} \\ & & \sigma_{22} \end{pmatrix} \right].$$

Clearly, under the data structure in Table 1, the parameter σ_{12} is not estimable from the samples. The conditional independence assumption in (1) implies that $\sigma_{12} = \sigma_{1x}\sigma_{2x}/\sigma_{xx}$ and $\rho_{12} = \rho_{1x}\rho_{2x}$. That is, σ_{12} is completely determined from other parameters, rather than estimated directly from the realized samples.

Synthetic data imputation under the conditional independence assumption in this case can be implemented in two steps:

[Step 1] Estimate $f(y_1 | x)$ from Sample A, and denote the estimate by $\hat{f}_a(y_1 | x)$.

[Step 2] For each element i in Sample B, use the x_i value to generate imputed value(s) of y_1 from $\hat{f}_a(y_1 | x_i)$.

Since y_1 values are never observed in Sample B, synthetic values of y_1 are created for all elements in Sample B, leading to synthetic imputation. Haziza (2009) provides a nice review of literature on imputation methodology. Kim & Rao (2012) present a model-assisted approach to synthetic imputation when only x is available in Sample B. Such synthetic imputation completely ignores the observed information in y_2 from Sample B.

Statistical matching based on conditional independence assumes that $Cov(y_1, y_2 | x) = 0$. Thus, the regression of y_2 on x and y_1 using the imputed data from the above synthetic imputation will estimate a zero regression coefficient for y_1 . That is, the estimate $\hat{\beta}_2$ for

$$\hat{y}_2 = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 y_1,$$

will estimate zero. Such analyses can be misleading if CI does not hold. To explain why, we

consider an omitted variable regression problem:

$$\begin{aligned} y_1 &= \beta_0^{(1)} + \beta_1^{(1)}x + \beta_2^{(1)}z + e_1 \\ y_2 &= \beta_0^{(2)} + \beta_1^{(2)}x + \beta_2^{(2)}z + e_2 \end{aligned}$$

where z, e_1, e_2 are independent and are not observed. Unless $\beta_2^{(1)} = \beta_2^{(2)} = 0$, the latent variable z is an unobservable confounding factor that explains why $Cov(y_1, y_2 | x) \neq 0$. Thus, the coefficient on y_1 in the population regression of y_2 on x and y_1 is not zero,

We consider an alternative approach which is not built on the conditional independence assumption. First, assume that we can decompose x as $x = (x_1, x_2)$ such that

$$\begin{aligned} (i) \quad & f(y_2 | x_1, x_2, y_1) = f(y_2 | x_1, y_1) \\ (ii) \quad & f(y_1 | x_1, x_2 = a) \neq f(y_1 | x_1, x_2 = b) \end{aligned}$$

for some $a \neq b$. Thus, x_2 is conditionally independent of y_2 given x_1 and y_1 but x_2 is correlated with y_1 given x_1 . Note that x_1 may be null or have a degenerate distribution, such as an intercept. The variable x_2 satisfying the above two conditions is often called an instrumental variable (IV) for y_1 . The directed acyclic graph in Figure 1 illustrates the dependence structure of a model with an instrumental variable. Ridder & Moffit (2007) used “exclusion restrictions” to describe the instrumental variable assumption. One example where the instrumental variable assumption is reasonable is repeated surveys. In the repeated survey, suppose that y_t is the study variable at year t and satisfies Markov property

$$P(y_{t+1} | y_1, \dots, y_t) = P(y_{t+1} | y_t),$$

where $P(y_t)$ denotes a cumulative distribution function. In this case, y_{t-1} is an instrumental variable for y_t . In fact, any last observation of $y_s (s \leq t)$ is the instrumental variable for y_t .

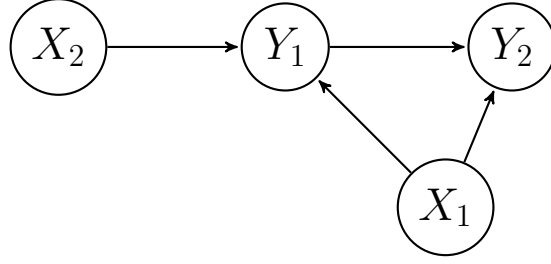


Figure 1: Graphical illustration of the dependence structure for a model in which x_2 is an instrumental variable for y_1 and x_1 is an additional covariate in the models for y_2 and y_1 .

Under the instrumental variable assumption, one can use two-step regression to estimate the regression parameters of a linear model. The following example presents the basic ideas.

Example 2.1. Consider the two sample data structure in Table 1. We assume the following linear regression model:

$$y_{2i} = \beta_0 + \beta_1 x_{1i} + \beta_2 y_{1i} + e_i, \quad (2)$$

where $e_i \sim (0, \sigma_e^2)$ and e_i is independent of (x_{1j}, x_{2j}, y_{1j}) for all i, j . In this case, a consistent estimator of $\beta = (\beta_0, \beta_1, \beta_2)$ can be obtained by the two-stage least squares (2SLS) method as follows:

1. From sample A , fit the following “working model” for y_1

$$y_{1i} = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + u_i, \quad u_i \sim (0, \sigma_u^2) \quad (3)$$

to obtain a consistent estimator of $\alpha = (\alpha_0, \alpha_1, \alpha_2)'$ defined by

$$\hat{\alpha} = (\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2)' = (X'X)^{-1} X'Y_1$$

where $X = [X_0, X_1, X_2]$ is a matrix whose i -th row is $(1, x_{1i}, x_{2i})$ and Y_1 is a vector with y_{1i} being the i -th component.

2. A consistent estimator of $\beta = (\beta_0, \beta_1, \beta_2)'$ is obtained by the least squares method for the regression of y_{2i} on $(1, x_{1i}, \hat{y}_{1i})$ where $\hat{y}_{1i} = \hat{\alpha}_0 + \hat{\alpha}_1 x_{1i} + \hat{\alpha}_2 x_{2i}$.

Asymptotic unbiasedness of the 2SLS estimator under the instrumental variable assumption is discussed in Appendix A. The 2SLS method is not directly applicable if the regression model (2) is nonlinear. Also, while the 2SLS method gives estimates of the regression parameters, 2SLS does not provide consistent estimators for more general parameters such as $\theta = Pr(y_2 < 1 \mid y_1 < 3)$. Stochastic imputation can provide a solution for estimating a more general class of parameters. We explain how to modify parametric fractional imputation of Kim (2011) to address general purpose estimation in statistical matching problems.

3 Fractional imputation

We now describe the fractional imputation methods for statistical matching without using the CI assumption. The use of fractional imputation for statistical matching was originally presented in Chapter 9 of Kim and Shao (2013). To explain the idea, note that y_1 is missing in Sample B and our goal is to generate y_1 from the conditional distribution of y_1 given the observations. That is, we wish to generate y_1 from

$$f(y_1 \mid x, y_2) \propto f(y_2 \mid x, y_1) f(y_1 \mid x). \quad (4)$$

To satisfy model identifiability, we may assume that x_2 is an IV for y_1 . Under IV assumption, (4) reduces to

$$f(y_1 \mid x, y_2) \propto f(y_2 \mid x_1, y_1) f(y_1 \mid x).$$

To generate y_1 from (4), we can consider the following two-step imputation:

1. Generate y_1^* from $\hat{f}_a(y_1 \mid x)$.
2. Accept y_1^* if $f(y_2 \mid x, y_1^*)$ is sufficiently large.

Note that the first step is the usual method under the conditional independence assumption. The second step incorporates the information in y_2 . The determination of whether $f(y_2 \mid x, y_1^*)$ is sufficiently large required for Step 2 is often made by applying a Markov Chain Monte Carlo (MCMC)

method such as the Metropolis-Hastings algorithm (Chib & Greenberg, 1995). That is, let $y_1^{(t-1)}$ be the current value of y_1 in the Markov Chain. Then, we accept y_1^* with probability

$$R(y_1^*, y_1^{(t-1)}) = \min \left\{ 1, \frac{f(y_2 | x, y_1^*)}{f(y_2 | x, y_1^{(t-1)})} \right\}.$$

Such algorithms can be computationally cumbersome because of slow convergence of the MCMC algorithm.

Parametric fractional imputation of Kim (2011) enables generating imputed values in (4) without requiring MCMC. The following EM algorithm by fractional imputation can be used:

1. For each $i \in B$, generate m imputed values of y_{1i} , denoted by $y_{1i}^{*(1)}, \dots, y_{1i}^{*(m)}$, from $\hat{f}_a(y_1 | x_i)$, where $\hat{f}_a(y_1 | x)$ denotes the estimated density for the conditional distribution of y_1 given x obtained from sample A.
2. Let $\hat{\theta}_t$ be the current parameter value of θ in $f(y_2 | x, y_1)$. For the j -th imputed value $y_{1i}^{*(j)}$, assign the fractional weight

$$w_{ij(t)}^* \propto f(y_{2i} | x_i, y_{1i}^{*(j)}; \hat{\theta}_t)$$

such that $\sum_{j=1}^m w_{ij}^* = 1$.

3. Solve the fractionally imputed score equation for θ

$$\sum_{i \in B} w_{ib} \sum_{j=1}^m w_{ij(t)}^* S(\theta; x_i, y_{1i}^{*(j)}, y_{2i}) = 0 \quad (5)$$

to obtain $\hat{\theta}_{t+1}$, where $S(\theta; x, y_1, y_2) = \partial \log f(y_2 | x, y_1; \theta) / \partial \theta$, and w_{ib} is the sampling weight of unit i in Sample B.

4. Go to step 2 and continue until convergence.

In (5), note that, for sufficiently large m ,

$$\begin{aligned} \sum_{j=1}^m w_{ij(t)}^* S(\theta; x_i, y_{1i}^{*(j)}, y_{2i}) &\cong \frac{\int S(\theta; x_i, y_1, y_{2i}) f(y_{2i} | x_i, y_{1i}^{*(j)}; \hat{\theta}_t) \hat{f}_a(y_1 | x_i) dy_1}{\int f(y_{2i} | x_i, y_{1i}^{*(j)}; \hat{\theta}_t) \hat{f}_a(y_1 | x_i) dy_1} \\ &= E \left\{ S(\theta; x_i, Y_1, y_{2i}) | x_i, y_{2i}; \hat{\theta}_t \right\}. \end{aligned}$$

If y_{i1} is categorical, then the fractional weight can be constructed by the conditional probability corresponding to the realized imputed value (Ibrahim, 1990). Step 2 is used to incorporate observed information of y_{i2} in Sample B. Note that Step 1 is not repeated for each iteration. Only Step 2 and Step 3 are iterated until convergence. Because Step 1 is not iterated, convergence is guaranteed and the observed likelihood increases. See Theorem 2 of Kim (2011).

Remark 3.1. *In Section 2, we introduce IV only because this is what it is typically done in the literature to ensure identifiability. The proposed method itself does not rely on this assumption. To illustrate a situation where we can identify the model without introducing the IV assumption, suppose that the model is*

$$\begin{aligned} y_2 &= \beta_0 + \beta_1 x + \beta_2 y_1 + e_2 \\ y_1 &= \alpha_0 + \alpha_1 x + e_1 \end{aligned}$$

with $e_1 \sim N(0, x\sigma_1^2)$ and $e_2 \mid e_1 \sim N(0, \sigma_2^2)$, then

$$f(y_2 \mid x) = \int f(y_2 \mid x, y_1) f(y_1 \mid x) dy_1$$

is also a normal distribution with mean $(\beta_0 + \beta_2 \alpha_0) + (\beta_1 + \beta_2 \alpha_1)x$ and variance $\sigma_2^2 + \beta_2^2 \sigma_1^2 x$. Under the data structure in Table 1, such a model is identified without assuming the IV assumption.

Instead of generating $y_{1i}^{*(j)}$ from $\hat{f}_a(y_1 \mid x_i)$, we can consider a hot-deck fractional imputation (HDFI) method, where all the observed values of y_{1i} in Sample A are used as imputed values. In this case, the fractional weights in Step 2 are given by

$$w_{ij}^*(\hat{\theta}_t) \propto w_{ij0}^* f(y_{2i} \mid x_i, y_{1i}^{*(j)}; \hat{\theta}_t),$$

where

$$w_{ij0}^* = \frac{\hat{f}_a(y_{1j} \mid x_i)}{\sum_{k \in A} w_{ka} \hat{f}_a(y_{1j} \mid x_k)}. \quad (6)$$

The initial fractional weight w_{ij0}^* in (6) is computed by applying importance weighting with

$$\hat{f}_a(y_{1j}) = \int \hat{f}_a(y_{1j} \mid x) \hat{f}_a(x) dx \propto \sum_{i \in A} w_{ia} \hat{f}_a(y_{1j} \mid x_i)$$

as the proposal density for y_{1j} . The M-step is the same as for parametric fractional imputation. See Kim & Yang (2013) for more details on HDFI. In practice, we may use a single imputed value for each unit. In this case, the fractional weights can be used as the selection probability in Probability-Proportional-to-Size (PPS) sampling of size $m = 1$.

For variance estimation, we can either use a linearization method or a resampling method. We first consider variance estimation for the maximum likelihood estimator (MLE) of θ . If we use a parametric model $f(y_1 | x) = f(y_1 | x; \theta_1)$ and $f(y_2 | x, y_1; \theta_2)$, the MLE of $\theta = (\theta_1, \theta_2)$ is obtained by solving

$$[S_1(\theta_1), \bar{S}_2(\theta_1, \theta_2)] = (0, 0), \quad (7)$$

where $S_1(\theta_1) = \sum_{i \in A} w_{ia} S_{i1}(\theta_1)$, $S_{i1}(\theta_1) = \partial \log f(y_{1i} | x_i; \theta_1) / \partial \theta_1$ is the score function of θ_1 ,

$$\bar{S}_2(\theta_1, \theta_2) = E\{S_2(\theta_2) | X, Y_2; \theta_1, \theta_2\},$$

$S_2(\theta_2) = \sum_{i \in B} w_{ib} S_{i2}(\theta_2)$, and $S_{i2}(\theta_2) = \partial \log f(y_{2i} | x_i, y_{1i}; \theta_2) / \partial \theta_2$ is the score function of θ_2 .

Note that we can write $\bar{S}_2(\theta_1, \theta_2) = \sum_{i \in B} w_{ib} E\{S_{i2}(\theta_2) | x_i, y_{2i}; \theta\}$. Thus,

$$\begin{aligned} \frac{\partial}{\partial \theta'_1} \bar{S}_2(\theta) &= \sum_{i \in B} w_{ib} \frac{\partial}{\partial \theta'_1} \left[\frac{\int S_{i2}(\theta_2) f(y_1 | x_i; \theta_1) f(y_{2i} | x_i, y_1; \theta_2) dy_1}{\int f(y_1 | x_i; \theta_1) f(y_{2i} | x_i, y_1; \theta_2) dy_1} \right] \\ &= \sum_{i \in B} w_{ib} E\{S_{i2}(\theta_2) S_{i1}(\theta_1)' | x_i, y_{2i}; \theta\} \\ &\quad - \sum_{i \in B} w_{ib} E\{S_{i2}(\theta_2) | x_i, y_{2i}; \theta\} E\{S_{i1}(\theta_1)' | x_i, y_{2i}; \theta\} \end{aligned}$$

and

$$\begin{aligned} \frac{\partial}{\partial \theta'_2} \bar{S}_2(\theta) &= \sum_{i \in B} w_{ib} \frac{\partial}{\partial \theta'_2} \left[\frac{\int S_{i2}(\theta_2) f(y_1 | x_i; \theta_1) f(y_{2i} | x_i, y_1; \theta_2) dy_1}{\int f(y_1 | x_i; \theta_1) f(y_{2i} | x_i, y_1; \theta_2) dy_1} \right] \\ &= \sum_{i \in B} w_{ib} E\left\{ \frac{\partial}{\partial \theta'_2} S_{i2}(\theta_2) | x_i, y_{2i}; \theta \right\} \\ &\quad + \sum_{i \in B} w_{ib} E\{S_{i2}(\theta_2) S_{i2}(\theta_2)' | x_i, y_{2i}; \theta\} \\ &\quad - \sum_{i \in B} w_{ib} E\{S_{i2}(\theta_2) | x_i, y_{2i}; \theta\} E\{S_{i2}(\theta_2)' | x_i, y_{2i}; \theta\}. \end{aligned}$$

Now, $\partial \bar{S}_2(\theta)/\partial \theta'_1$ can be consistently estimated by

$$\hat{B}_{21} = \sum_{i \in B} w_{ib} \sum_{j=1}^m w_{ij}^* S_{2ij}^*(\hat{\theta}_2) \left\{ S_{1ij}^*(\hat{\theta}_1) - \bar{S}_{1i}^*(\hat{\theta}_1) \right\}', \quad (8)$$

where $S_{1ij}^*(\hat{\theta}_1) = S_1(\hat{\theta}_1; x_i, y_{1i}^{*(j)})$, $S_{2ij}^*(\hat{\theta}_2) = S_2(\hat{\theta}_2; x_i, y_{1i}^{*(j)}, y_{2i})$, and $\bar{S}_{1i}^*(\hat{\theta}_1) = \sum_{j=1}^m w_{ij}^* S_1(\hat{\theta}_1; x_i, y_{1i}^{*(j)})$.

Also, $\partial \bar{S}_2(\theta)/\partial \theta'_2$ can be consistently estimated by

$$-\hat{I}_{22} = \sum_{i \in B} w_{ib} \sum_{j=1}^m w_{ij}^* \dot{S}_{2ij}^*(\hat{\theta}_2) - \hat{B}_{22} \quad (9)$$

where

$$\hat{B}_{22} = \sum_{i \in B} w_{ib} \sum_{j=1}^m w_{ij}^* S_{2ij}^*(\hat{\theta}_2) \left\{ S_{2ij}^*(\hat{\theta}_2) - \bar{S}_{2i}^*(\hat{\theta}_2) \right\}',$$

$\dot{S}_{2ij}^*(\theta_2) = \partial S_2(\theta_2; x_i, y_{1i}^{*(j)}, y_{2i})/\partial \theta'_2$ and $\bar{S}_{2i}^*(\theta_2) = \sum_{j=1}^m w_{ij}^* S_{2ij}^*(\theta_2)$.

Using a Taylor expansion with respect to θ_1 ,

$$\begin{aligned} \bar{S}_2(\hat{\theta}_1, \theta_2) &\cong \bar{S}_2(\theta_1, \theta_2) - E \left\{ \frac{\partial}{\partial \theta'_1} \bar{S}_2(\theta) \right\} \left[E \left\{ \frac{\partial}{\partial \theta'_1} S_1(\theta_1) \right\} \right]^{-1} S_1(\theta_1) \\ &= \bar{S}_2(\theta) + K S_1(\theta_1), \end{aligned}$$

and we can write

$$V(\hat{\theta}_2) \doteq \left\{ E \left(\frac{\partial}{\partial \theta'_2} \bar{S}_2 \right) \right\}^{-1} V \{ \bar{S}_2(\theta) + K S_1(\theta_1) \} \left\{ E \left(\frac{\partial}{\partial \theta'_2} \bar{S}_2 \right) \right\}^{-1'}.$$

Writing

$$\bar{S}_2(\theta) = \sum_{i \in B} w_{ib} \bar{s}_{2i}(\theta),$$

with $\bar{s}_{2i}(\theta) = E\{S_{i2}(\theta_2) \mid x_i, y_{2i}; \theta\}$, a consistent estimator of $V \{ \bar{S}_2(\theta) \}$ can be obtained by applying a design-consistent variance estimator to $\sum_{i \in B} w_{ib} \hat{s}_{2i}$ with $\hat{s}_{2i} = \sum_{j=1}^m w_{ij}^* S_{2ij}^*(\hat{\theta}_2)$. Under simple random sampling for Sample B, we have

$$\hat{V} \{ \bar{S}_2(\theta) \} = n_B^{-2} \sum_{i \in B} \hat{s}_{2i} \hat{s}'_{2i}.$$

Also,

$$V \{K S_1(\theta_1)\}$$

is consistently estimated by

$$\hat{V}_2 = \hat{K} \hat{V}(S_1) \hat{K}',$$

where $\hat{K} = \hat{B}_{21} \hat{I}_{11}^{-1}$, \hat{B}_{21} is defined in (8), and $\hat{I}_{11} = -\partial S_1(\theta_1)/\partial \theta_1'$ evaluated at $\theta_1 = \hat{\theta}_1$. Since the two terms $\bar{S}_2(\theta)$ and $S_1(\theta_1)$ are independent, the variance can be estimated by

$$\hat{V}(\hat{\theta}) \doteq \hat{I}_{22}^{-1} \left[\hat{V} \{ \bar{S}_2(\theta) \} + \hat{V}_2 \right] \hat{I}_{22}^{-1'},$$

where \hat{I}_{22} is defined in (9).

More generally, one may consider estimation of a parameter η defined as a root of the census estimating equation $\sum_{i=1}^N U(\eta; x_i, y_{1i}, y_{2i}) = 0$. Variance estimation of the FI estimator of η computed from $\sum_{i \in B} w_{ib} \sum_{j=1}^m w_{ij}^* U(\eta; x_i, y_{1i}^{*(j)}, y_{2i}) = 0$ is discussed in Appendix B.

4 Split questionnaire survey design

In Section 3, we consider the situation where Sample A and Sample B are two independent samples from the same target population. We now consider another situation of a split questionnaire design where the original sample S is selected from a target population and then Sample A and Sample B are randomly chosen such that $A \cup B = S$ and $A \cap B = \phi$. We observe (x, y_1) from Sample A and observe (x, y_2) from Sample B. We are interested in creating fully augmented data with observation (x, y_1, y_2) in S .

Such split questionnaire survey designs are gaining popularity because they reduce response burden (Ragunathan & Grizzle, 1995; Chipperfield & Steel, 2009). Split questionnaire designs have been investigated for the Consumer Expenditure survey (Gonzalez & Eltinge, 2008) and the National Assessment of Educational Progress (NAEP) survey in the US. In applications of split-questionnaire designs, analysts may be interested in multiple parameters such as the mean of y_1 and the mean of y_2 , in addition to the coefficient in the regression of y_2 on y_1 .

To construct a fully augmented dataset in S , we still assume the instrumental variable assumption given in (i) and (ii) of Section 2. That is, we assume $x = (x_1, x_2)$, where x_2 satisfies $f(y_2 | x_1, x_2, y_1) = f(y_2 | x_1, y_1)$ and $f(y_1 | x_1, x_2 = a) \neq f(y_1 | x_1, x_2 = b)$ for some $a \neq b$. One can use the sample data for inference about the marginal distribution of y_1 , the marginal distribution of y_2 , and the conditional distribution of y_1 or y_2 given x . The instrumental variable assumption permits identification of the parameters defining the joint distribution of y_1 and y_2 . Estimators of parameters in the marginal distributions of y_1 and y_2 based on the fully imputed data set are more efficient than estimators based only on the sample data if y_1 and y_2 are correlated.

In some split questionnaire designs (i.e. Raghunathan & Grizzle (1995)), the sample design is constructed so that every pair of questions is assigned to some subsample. This restriction on the design permits inference for joint distributions. The instrumental variable assumption allows inference for joint distributions with more general designs where some pairs of questions (i.e., questions leading to responses y_2 and y_1) are never asked to the same individual.

We consider a design where the original Sample S is partitioned into two subsamples: A and B . We assume that x_i is observed for $i \in S$, y_{1i} is collected for $i \in A$ and y_{2i} is collected for $i \in B$. (For simplicity, we assume that no nonresponse occurs for either Sample A or Sample B .) The probability of selection into A or B may depend on x_i but can not depend on y_{1i} or y_{2i} . As a consequence, the design used to select subsample A or B is non-informative for the specified model (Fuller, 2009, Chapter 6). We let w_i denote the sampling weight associated with the full sample S . We assume a procedure is available for estimating the variance of an estimator of the form $\hat{Y} = \sum_{i \in S} w_i y_i$, and we denote the variance estimator by $\hat{V}_s(\sum_{i \in S} w_i y_i)$.

A procedure for obtaining a fully imputed data set is as follows. First, use the procedure of Section 3 to obtain imputed values $\{y_{1i}^{*(j)} : i \in B, j = 1, \dots, m\}$ and an estimate, $\hat{\theta}$, of the parameter in the distribution $f(y_2 | y_1, x_1; \theta)$. The estimate $\hat{\theta}$ is obtained by solving

$$\sum_{i \in B} w_i \sum_{j=1}^m w_{ij}^* S_2(\theta; x_{1i}, y_{1i}^{*(j)}, y_{2i}) = 0, \quad (10)$$

where $S_2(\theta; x_1, y_1, y_2) = \partial \log f(y_2 | y_1, x_1; \theta) / \partial \theta$. Given $\hat{\theta}$, generate imputed values $y_{2i}^{*(j)} \sim$

$f(y_2 | y_{1i}, x_{1i}; \hat{\theta})$, for $i \in A$ and $j = 1, \dots, m$.

Under the instrumental variable assumption, the parameter estimator $\hat{\theta}$ generated by solving (10) is fully efficient in the sense that the imputed value of y_{2i} for Sample A leads to no efficiency gain. To see this, note that the score equation using the imputed value of y_{2i} is computed by

$$\sum_{i \in A} w_i m^{-1} \sum_{j=1}^m S_2(\theta; x_{1i}, y_{1i}, y_{2i}^{*(j)}) + \sum_{i \in B} w_i m^{-1} \sum_{j=1}^m w_{ij}^* S_2(\theta; x_{1i}, y_{1i}^*, y_{2i}) = 0. \quad (11)$$

Because $y_{2i}^{*(1)}, \dots, y_{2i}^{*(m)}$ are generated from $f(y_2 | y_{1i}, x_{1i}; \hat{\theta})$,

$$p \lim_{m \rightarrow \infty} \sum_{i \in A} w_i m^{-1} \sum_{j=1}^m S_2(\theta; x_{1i}, y_{1i}, y_{2i}^{*(j)}) = \sum_{i \in A} w_i E\{S_2(\theta; x_{1i}, y_{1i}, Y_2) | y_{1i}, x_{1i}; \hat{\theta}\}.$$

Thus, by the property of score function, the first term of (11) evaluated at $\theta = \hat{\theta}$ is close to zero and the solution to (11) is essentially the same as the solution to (10). That is, there is no efficiency gain in using the imputed value of y_{2i} in computing the MLE for θ in $f(y_2 | y_1, x_1; \theta)$.

However, the imputed values of y_{2i} can improve the efficiency of inferences for parameters in the joint distribution of (y_{1i}, y_{2i}) . As a simple example, consider estimation of μ_2 , the marginal mean of y_{2i} . Under simple random sampling, the imputed estimator of $\theta = \mu_2$ is

$$\hat{\theta}_{I,m} = \frac{1}{n} \left\{ \sum_{i \in A} \left(m^{-1} \sum_{j=1}^m y_{2i}^{*(j)} \right) + \sum_{i \in B} y_{2i} \right\}. \quad (12)$$

For sufficiently large m , we can write

$$\begin{aligned} \hat{\theta}_{I,m} &= \frac{1}{n} \left\{ \sum_{i \in A} \hat{y}_{2i} + \sum_{i \in B} y_{2i} \right\} \\ &= \frac{1}{n} \left\{ \sum_{i \in A} (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 y_{1i}) + \sum_{i \in B} y_{2i} \right\}, \end{aligned}$$

where $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ satisfies

$$\sum_{i \in B} (y_{2i} - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 y_{1i}) = 0$$

and $\hat{y}_{1i} = \hat{\alpha}_0 + \hat{\alpha}_1 x_{1i} + \hat{\alpha}_2 x_{2i}$ with $(\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2)$ satisfying

$$\sum_{i \in A} (y_{1i} - \hat{\alpha}_0 - \hat{\alpha}_1 x_{1i} - \hat{\alpha}_2 x_{2i}) = 0.$$

Under the regression model

$$y_{2i} = \beta_0 + \beta_1 x_{1i} + \beta_2 \hat{y}_{1i} + e_i$$

where $e_i \sim (0, \sigma_e^2)$, the variance of $\hat{\theta}_{I,m}$ is, for sufficiently large m ,

$$V(\hat{\theta}_{I,m}) = \frac{1}{n} V(y_2) + \left(\frac{1}{n_b} - \frac{1}{n} \right) V(e)$$

which is smaller than the variance of the direct estimator $\hat{\theta} = n_b^{-1} \sum_{i \in B} y_{2i}$.

5 Measurement error models

We now consider an application of statistical matching to the problem of measurement error models. Suppose that we are interested in the parameter θ in the conditional distribution $f(y \mid x; \theta)$. In the original sample, instead of observing (x_i, y_i) , we observe (z_i, y_i) , where z_i is a contaminated version of x_i . Because inference for θ based on (z_i, y_i) may be biased, additional information is needed. One common way to obtain additional information is to collect (x_i, z_i) in an external calibration study. In this case, we observe (x_i, z_i) in Sample A and (z_i, y_i) in Sample B, where sample A is the calibration sample, and Sample B is the main sample. Guo & Little (2011) discuss an application of external calibration.

The external calibration framework can be expressed as a statistical matching problem. Table 2 makes the connection between statistical matching and external calibration explicit. The (x_i, z_i, y_i) in the measurement error framework correspond to the (y_{1i}, x_{2i}, y_{2i}) in the setting of statistical matching. A straightforward extension of the measurement error model considered here incorporates additional covariates, such as the x_{1i} of the statistical matching framework.

	z_i	x_i	y_i
Survey A (calibration study)	o	o	
Survey B (main study)	o		o

Table 2: Data structure for measurement error model

An instrumental variable assumption permits inference for θ based on data with the structure of Table 1. In the notation of the measurement error model, the instrumental variable assumption is

$$f(y_i | x_i, z_i) = f(y_i | x_i) \text{ and } f(z_i | x_i = a) \neq f(z_i | x_i = b), \quad (13)$$

for some $a \neq b$. The instrumental variable assumption may be judged reasonable in applications related to error in covariates because the subject-matter model of interest is $f(y_i | x_i)$, and z_i is a contaminated version of x_i that contains no additional information about y_i given x_i .

For fully parametric $f(y_i | x_i)$, $f(z_i | x_i)$ and $f(x_i)$, one can use parametric fractional imputation to execute the EM algorithm. This method requires evaluating the conditional expectation of the complete-data score function given the observed values. To evaluate the conditional expectation using fractional imputation, we first express the conditional distribution of x given (z, y) as,

$$f(x | z, y) \propto f(y | x) f(x | z). \quad (14)$$

We let an estimator $\hat{f}_a(x_i | z_i)$ of $f(x_i | z_i)$ be available from the calibration sample (Sample A). Implementation of the EM algorithm via fractional imputation involves the following steps:

1. For each $i \in B$, generate $x_i^{*(j)}$ from $\hat{f}_a(x | z_i)$, for $j = 1, \dots, m$,
2. Compute the fractional weights

$$w_{ij}^* \propto f(y_i | x_i^{*(j)}; \hat{\theta}_t).$$

3. Update θ by solving

$$\sum_{i \in B} w_{ib} \sum_{j=1}^m w_{ij}^* S(\theta; x_i^{*(j)}, y_i) = 0,$$

where $S(\theta; x_i^{*(j)}, y_i) = \partial \log\{f(y | x; \theta)\} / \partial \theta$.

4. Go to Step 2 until convergence.

The method above requires generating data from $f(x | z)$. For some nonlinear models or models with non-constant variances, simulating from the conditional distribution of x given z may require Monte Carlo methods such as accept-reject or Metropolis Hastings. The simulation of Section 6.2 exemplifies a simulation in which the conditional distribution of $x | z$ has no closed form expression. In this case, we may consider an alternative approach, which may be computationally simpler. To describe this approach, let $h(x | z)$ be the “working” conditional distribution, such as the normal distribution, from which samples are easily generated. A special case of $h(x | z)$ is $f(x)$, the marginal density of X , which is used for selecting donors for HDFI. We assume that estimates $\hat{f}_a(x | z)$ and $\hat{h}_a(x | z)$ of $f(x | z)$ and $h(x | z)$, respectively, are available from Sample A. Implementation of the EM algorithm via fractional imputation then involves the following steps:

1. For each $i \in B$, generate $x_i^{*(j)}$ from $\hat{h}_a(x | z_i)$, for $j = 1, \dots, m$,
2. Compute the fractional weights

$$w_{ij}^* \propto f(y_i | x_i^{*(j)}; \hat{\theta}_t) \hat{f}_a(x_i^{*(j)} | z_i) / \hat{h}_a(x_i^{*(j)} | z_i). \quad (15)$$

3. Update θ by solving

$$\sum_{i \in B} w_{ib} \sum_{j=1}^m w_{ij}^* S(\theta; x_i^{*(j)}, y_i) = 0.$$

4. Go to Step 2 until convergence.

Remark 5.1. *Variance estimation is a straightforward application of the linearization method in Section 3. The hot-deck fractional imputation method described in Section 3 with fractional*

weights defined in (6) also extends readily to the measurement error setting. For HDFI, the proposal distribution $\hat{h}_a(x \mid z)$ can be the empirical distribution with weights proportional to the sampling weights in Sample A. The imputed values are the n_A values of x_i . The weight w_{ij}^* used for HDFI is

$$w_{ij}^* \propto f(y_i \mid x_i^{*(j)}; \hat{\theta}_t) \hat{f}_a(x_i^{*(j)} \mid z_i) / w_{ja}, \quad (16)$$

where $x_i^{*(j)} = x_j$ from sample A, and w_{ja} is the associated sampling weight.

6 Simulation study

To test our theory, we present two limited simulation studies. The first simulation study considers the setup of combining two independent surveys of partial observation to obtain joint analysis. The second simulation study considers the setup of measurement error models with external calibration.

6.1 Simulation One

To compare the proposed methods with the existing methods, we generate 5,000 Monte Carlo samples of (x_i, y_{1i}, y_{2i}) with size $n = 400$, where

$$\begin{pmatrix} y_{1i} \\ x_i \end{pmatrix} \sim N \left(\begin{bmatrix} 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 1 & 0.7 \\ 0.7 & 1 \end{bmatrix} \right), \quad (17)$$

$$y_{2i} = \beta_0 + \beta_1 y_{1i} + e_i,$$

$e_i \sim N(0, \sigma^2)$, and $\beta = (\beta_0, \beta_1, \sigma^2) = (1, 1, 1)$. Note that, in this setup, we have $f(y_2 \mid x, y_1) = f(y_2 \mid y_1)$ and so the variable x plays the role of the instrumental variable for y_1 .

Instead of observing (x_i, y_{1i}, y_{2i}) jointly, we assume that only (y_1, x) are observed in Sample A and only (y_2, x) are observed in Sample B, where Sample A is obtained by taking the first $n_a = 400$ elements and Sample B is obtained by taking the remaining $n_b = 400$ elements from the original sample. We are interested in estimating four parameters: three regression parameters $\beta_0, \beta_1, \sigma^2$ and $\pi = P(y_1 < 2, y_2 < 3)$, the proportion of $y_1 < 2$ and $y_2 < 3$. Four methods are considered in estimating the parameters:

1. Full sample estimation (Full): Uses the complete observation of (y_{1i}, y_{2i}) in Sample B.
2. Stochastic regression imputation (SRI): Use the regression of y_1 on x from Sample A to obtain $(\hat{\alpha}_0, \hat{\alpha}_1, \hat{\sigma}_1^2)$, where the regression model is $y_1 = \alpha_0 + \alpha_1 x + e_1$ with $e_1 \sim (0, \sigma_1^2)$. For each $i \in B$, $m = 10$ imputed values are generated by $y_{1i}^{*(j)} = \hat{\alpha}_0 + \hat{\alpha}_1 x_i + e_i^{*(j)}$ where $e_i^{*(j)} \sim N(0, \hat{\sigma}_1^2)$.
3. Parametric fractional imputation (PFI) with $m = 10$ using the instrumental variable assumption.
4. Hot-deck fractional imputation (HDFI) with $m = 10$ using the instrumental variable assumption.

Table 3 presents Monte Carlo means and Monte Carlo variances of the point estimators of the four parameters of interest. SRI shows large biases for all parameters considered because it is based on the conditional independence assumption. Both PFI and HDFI provide nearly unbiased estimators for all parameters. Estimators from HDFI are slightly more efficient than those from PFI because the two-step procedure in HDFI uses the full set of respondents in the first step. The theoretical asymptotic variance of $\hat{\beta}_1$ computed from PFI is

$$V(\hat{\beta}_1) \doteq \frac{1}{(0.7)^2} \frac{1}{400} 2 \left(1 - \frac{0.7^2}{2} \right) + \frac{1}{(0.7)^2} \frac{1}{400} (1 - 0.7^2) \doteq 0.0103$$

which is consistent with the simulation result in Table 3. In addition to point estimation, we also compute variance estimators for PFI and HDFI methods. Variance estimators show small relative biases (less than 5% in absolute values) for all parameters. Variance estimation results are not presented here for brevity.

The proposed method is based on the instrumental variable assumption. To study the sensitivity of the proposed fractional imputation method, we performed an additional simulation study. Now, instead of generating y_{2i} from (17), we use

$$y_{2i} = 0.5 + y_{1i} + \rho(x_i - 3) + e_i, \tag{18}$$

Parameter	Method	Mean	Variance
β_0	Full	1.00	0.0123
	SRI	1.90	0.0869
	PFI	1.00	0.0472
	HDFI	1.00	0.0465
β_1	Full	1.00	0.00249
	SRI	0.54	0.01648
	PFI	1.00	0.01031
	HDFI	1.00	0.01026
σ^2	Full	1.00	0.00482
	SRI	1.73	0.01657
	PFI	0.99	0.02411
	HDFI	0.99	0.02270
π	Full	0.374	0.00058
	SRI	0.305	0.00255
	PFI	0.375	0.00059
	HDFI	0.375	0.00057

Table 3: Monte Carlo means and variances of point estimators from Simulation One. (SRI, stochastic regression imputation; PFI, parametric fractional imputation; HDFI; hot-deck fractional imputation)

where $e_i \sim N(0, 1)$ and ρ can take non-zero values. We use three values of ρ , $\rho \in \{0, 0.1, 0.2\}$, in the sensitivity analysis and apply the same PFI and HDFI procedure that is based on the assumption that x is an instrumental variable for y_1 . Such assumption is satisfied for $\rho = 0$, but it is weakly violated for $\rho = 0.1$ or $\rho = 0.2$. Using the fractionally imputed data in sample B, we estimated three parameters, $\theta_1 = E(Y_1)$, θ_2 is the slope for the simple regression of y_2 on y_1 , and $\theta_3 = P(y_1 < 2, y_2 < 3)$, the proportion of $y_1 < 2$ and $y_2 < 3$. Table 4 presents Monte Carlo means and variances of the point estimators for three parameters under three different models. In Table 4, the absolute values of the difference between the fractionally imputed estimator and the full sample estimator increase as the value of ρ increases, which is expected as the instrumental variable assumption is more severely violated for larger values of ρ , but the differences are relatively small for all cases. In particular, the estimator of θ_1 is not affected by the departure from the instrumental variable assumption. This is because the imputation estimator under incorrect imputation model still provides unbiased estimator for the population mean as long as the regression imputation model contains an intercept term (Kim & Rao, 2012). Thus, this limited sensitivity analysis suggests that the proposed method seems to provide comparable estimates when the instrumental variable assumption is weakly violated.

6.2 Simulation Two

In the second simulation study, we consider a binary response variable y_i , where

$$y_i \sim \text{Bernoulli}(p_i), \quad (19)$$

$$\text{logit}(p_i) = \gamma_0 + \gamma_x x_i,$$

and $x_i \sim N(\mu_x, \sigma_x^2)$. In the main sample, denoted by B , instead of observing (x_i, y_i) , we observe (z_i, y_i) , where

$$z_i = \beta_0 + \beta_1 x_i + u_i, \quad (20)$$

and $u_i \sim N(0, \sigma^2 | x_i |^{2\alpha})$. We observe (x_i, z_i) , $i = 1, \dots, n_A$ in a calibration sample, denoted by A. For the simulation, $n_A = n_B = 800$, $\gamma_0 = 1$, $\gamma_x = 1$, $\beta_0 = 0$, $\beta_1 = 1$, $\sigma^2 = 0.25$, $\alpha = 0.4$,

Model	Parameter	Method	Mean	Variance
$\rho = 0$	θ_1	Full	2.00	0.00235
		PFI	2.00	0.00352
		FHDI	2.00	0.00249
	θ_2	Full	1.00	0.00249
		PFI	1.00	0.01031
		FHDI	1.00	0.01026
	θ_3	Full	0.43	0.00061
		PFI	0.43	0.00059
		FHDI	0.43	0.00057
$\rho = 0.1$	θ_1	Full	2.00	0.00235
		PFI	2.00	0.00353
		FHDI	02.00	0.00250
	θ_2	Full	1.07	0.00248
		PFI	1.14	0.01091
		FHDI	1.14	0.01081
	θ_3	Full	0.44	0.00061
		PFI	0.45	0.00062
		FHDI	0.45	0.00059
$\rho = 0.2$	θ_1	Full	2.00	0.00235
		PFI	2.00	0.00353
		FHDI	2.00	0.00250
	θ_2	Full	1.14	0.00250
		PFI	1.28	0.01115
		FHDI	1.28	0.01102
	θ_3	Full	0.44	0.00061
		PFI	0.46	0.00066
		FHDI	0.46	0.00062

Table 4: Monte Carlo means and Monte Carlo variances of the two point estimators for sensitivity analysis in Simulation One (Full, full sample estimator; PFI, parametric fractional imputation; HDI; hot-deck fractional imputation)

$\mu_x = 0$, and $\sigma_x^2 = 1$. Primary interest is in estimation of γ_x and testing the null hypothesis that $\gamma_x = 1$. The MC sample size is 1000.

We compare the PFI and HDFI estimators of γ_x to three other estimators. Because the conditional distribution of x_i given z_i is non-standard, we use the weights of (15) and (16) to implement PFI and HDFI, where the proposal distribution $\hat{h}_a(x_i, | z_i)$ is an estimate of the marginal distribution of x_i based on the data from sample A. We consider the following five estimators:

1. *PFI*: For PFI, the proposal distribution for generating $x_i^{*(j)}$ is a normal distribution with mean $\hat{\mu}_x$ and variance $\hat{\sigma}_x^2$, where $\hat{\mu}_x$ and variance $\hat{\sigma}_x^2$ are the maximum likelihood estimates of μ_x and σ_x^2 based sample A. The fractional weight defined in (15) has the form,

$$w_{ij}^* \propto p_i^{y_i} (1 - p_i)^{1-y_i} \hat{f}_a(z_i | x_i), \quad (21)$$

where p_i is the function of (γ_0, γ_x) defined by (19), and $\hat{f}_a(z_i | x_i)$ is the estimate of $f(z_i | x_i)$ based on maximum likelihood estimation with the sample A data. The imputation size $m = 800$.

2. *HDFI*: For HDFI, instead of generating $x_i^{*(j)}$ from a normal distribution, the $\{x_i^{*(j)} : j = 1, \dots, 800\}$ are the 800 values of x_i from sample A.
3. *Naive*: A *naive* estimator is the estimator of the slope in the logistic regression of y_i on z_i for $i \in B$.
4. *Bayes*: We use the approach of Guo & Little (2011) to define a Bayes estimator. The model for this simulation differs from the model of Guo & Little (2011) in that the response of interest is binary. We implement GIBBS sampling with JAGS (Plummer, 2003), specifying diffuse proper prior distributions for the parameters of the model. Letting

$$\theta_1 = (\log(\sigma_x^2), \log(\sigma^2), \mu_x, \beta_0, \beta_1, \gamma_0, \gamma_x),$$

we assume a priori that $\theta_1 \sim N(0, 10^6 I_7)$, where I_7 is a 7×7 identity matrix, and the notation $N(0, V)$ denotes a normal distribution with mean 0 and covariance matrix V . The prior distribution for the power α is uniform on the interval $[-5, 5]$.

To evaluate convergence, we examine trace plots and potential scale reduction factors defined in Gelman et al. (2003) for 10 preliminary simulated data sets. We initiate three MCMC chains, each of length 1500 from random starting values and discard the first 500 iterations as burn-in. The potential scale reduction factors across the 10 simulated data sets range from 1.0 to 1.1, and the trace plots indicate that the chains mix well. To reduce computing time, we use 1000 iterations of a single chain for the main simulation, after discarding the first 500 for burn-in.

5. *A Weighted Regression Calibration (WRC) estimator.* The WRC estimator is a modification of the weighted regression calibration estimator defined in Guo & Little (2011) for a binary response variable. The computation for the weighted regression calibration estimator involves the following steps:

- (i) Using OLS, regress x_i on z_i for the calibration sample.
- (ii) Regress the logarithm of the squared residuals from step (i) on the logarithm of z_i^2 for the calibration sample. Let $\hat{\lambda}$ denote the estimated slope from the regression.
- (iii) Using WLS with weight $|z_i|^{2\hat{\lambda}}$, regress x_i on z_i for the calibration sample. Let $\hat{\eta}_0$ and $\hat{\eta}_1$ be the estimated intercept and slope, respectively, from the WLS regression.
- (iv) For each unit i in the main sample, let $\hat{x}_i = \hat{\eta}_0 + \hat{\eta}_1 z_i$.
- (v) The estimate of (γ_0, γ_x) is obtained from the logistic regression of y_i on \hat{x}_i for i in the main sample.

Table 5 contains the MC bias, variance, and MSE of the five estimators of γ_x . The naive estimator has a negative bias because z_i is a contaminated version of x_i . The variance of the PFI

Method	MC Bias	MC Variance	MC MSE
PFI	0.0239	0.0386	0.0392
HDFI	0.0246	0.0387	0.0393
Naive	-0.2241	0.0239	0.0742
Bayes	0.0406	0.0415	0.0432
WRC	0.112	0.0499	0.0625

Table 5: Monte Carlo (MC) means, variances, and mean squared errors (MSE) of point estimators of γ_x from Simulation Two. (PFI, parametric fractional imputation; HDFI, hot-deck fractional imputation; WRC, weighted regression calibration; MC, Monte Carlo; MSE, mean squared error)

estimator is modestly smaller than the variance of the HDFI estimator because the PFI estimator incorporates extra information through the parametric assumption about the distribution of x_i . The PFI and HDFI estimators are superior to the Bayes and WRC estimators.

We compute an estimate of the variance of the PFI and HDFI estimators of γ_x using the variance expression based on the linear approximation. We define the MC relative bias as the ratio of the difference between the MC mean of the variance estimator and the MC variance of the estimator to the MC variance of the estimator. The MC relative biases of the variance estimators for PFI and HDFI are -0.0096 and -0.0093, respectively.

7 Concluding Remarks

We approach statistical matching as a missing data problem and use PFI to obtain consistent estimators and corresponding variance estimators. The imputation approach applies more generally than two stage least squares, which is restricted to estimation of regression coefficients in linear models. Rather than rely on the often unrealistic conditional independence assumption, the imputation procedure derives from an assumption that an instrumental variable is available. The measurement error framework of Section 5 and Section 6.2, in which external calibration provides an independent measurement of the true covariate of interest, is a situation in which the study design

may be judged to support the instrumental variable assumption. Although the procedure is based on the instrumental variable assumption, the simulations of Section 6.1 show that the imputation method is robust to modest departures from the requirements of an instrumental variable.

The proposed methodology is applicable without the instrumental variable assumption, as long as the model is identified. If the model is not identifiable, then the EM algorithm for the proposed PFI method does not necessarily converge. In practice, one can treat the specified model identified if the EM sequence obtained from the specified model converges. The resulting analysis is consistent under the specified model. This is one of the main advantages of using the frequentist approach over Bayesian. In the Bayesian approach, it is possible to obtain the posterior values even under non-identified models and the resulting analysis can be misleading.

Statistical matching can also be used to evaluate effects of multiple treatments in observational studies. By properly applying statistical matching techniques, we can create an augmented data file of potential outcomes so that causal inference can be investigated with the augmented data file (Morgan & Winship, 2007). Such extensions will be presented elsewhere.

Acknowledgment

We thank Professor Yanyuan Ma, an anonymous referee and the AE for very constructive comments. The research of the first author was partially supported by a grant from NSF (MMS-121339) and Brain Pool program (131S-1-3-0476) from Korean Federation of Science and Technology Society. The research of the second author was supported by a Cooperative Agreement between the US Department of Agriculture Natural Resources Conservation Service and Iowa State University. The work of the third author was supported by the Bio-Synergy Research Project (2013M3A9C4078158) of the Ministry of Science, ICT and Future Planning through the National Research Foundation in Korea.

Appendix

A. Asymptotic unbiasedness of 2SLS estimator

Assume that we observe (y_1, x) in Sample A and observe (y_2, x) in Sample B. To be more rigorous, we can write (y_{1a}, x_a) to denote the observation (y_1, x) in Sample A. Also, we can write (y_{2b}, x_b) to denote the observations in Sample B. In this case, the model can be written as

$$\begin{aligned}y_{1a} &= \phi_0 1_a + \phi_1 x_{1a} + \phi_2 x_{2a} + e_{1a} \\y_{2b} &= \beta_0 1_b + \beta_1 x_{1b} + \beta_2 y_{1b} + e_{2b}\end{aligned}$$

with $E(e_{1a} \mid x_a) = 0$ and $E(e_{2b} \mid x_b, y_{1b}) = 0$. Note that y_{1b} is not observed from the sample. Instead, we use \hat{y}_{1b} using the OLS estimate obtained from Sample A.

Writing $X_a = [1_a, x_a]$ and $X_b = [1_b, x_b]$, we have $\hat{y}_{1b} = X_b(X_a'X_a)^{-1}X_a'y_{1a} = X_b\hat{\phi}_a$. The 2SLS estimator of $\beta = (\beta_0, \beta_1, \beta_2)'$ is then

$$\hat{\beta}_{2SLS} = (Z_b'Z_b)^{-1}Z_b'y_{2b}$$

where $Z_b = [1_b, x_{1b}, \hat{y}_{1b}]$. Thus, we have

$$\begin{aligned}\hat{\beta}_{2SLS} - \beta &= (Z_b'Z_b)^{-1}Z_b'(y_{2b} - Z_b\beta) \\&= (Z_b'Z_b)^{-1}Z_b'\{\beta_2(y_{1b} - \hat{y}_{1b}) + e_{2b}\}.\end{aligned}\tag{A.1}$$

We may write

$$y_{1b} = \phi_0 1_b + \phi_1 x_b + e_{1b} = X_b\phi + e_{1b}$$

where $E(e_{1b} \mid x_b) = 0$. Since

$$\begin{aligned}\hat{y}_{1b} &= X_b(X_a'X_a)^{-1}X_a'y_{1a} \\&= X_b(X_a'X_a)^{-1}X_a'(X_a\phi + e_{1a}) \\&= X_b\phi + X_b(X_a'X_a)^{-1}X_a'e_{1a},\end{aligned}$$

we have

$$y_{1b} - \hat{y}_{1b} = e_{1b} - X_b(X'_a X_a)^{-1} X'_a e_{1a}$$

and (A.1) becomes

$$\hat{\beta}_{2SLs} - \beta = (Z'_b Z_b)^{-1} Z'_b \{\beta_2 e_{1b} - \beta_2 X_b(X'_a X_a)^{-1} X'_a e_{1a} + e_{2b}\}. \quad (\text{A.2})$$

Assume that the two samples are independent. Thus, $E(e_{1b} \mid x_a, x_b, y_{1a}) = 0$. Also, $E\{(Z'_b Z_b)^{-1} Z'_b e_{2b} \mid x_a, x_b, y_{1a}, y_{1b}\} = 0$. Thus,

$$E\{\hat{\beta}_{2SLs} - \beta \mid x_a, x_b, y_{1a}\} = E\{-\beta_2 (Z'_b Z_b)^{-1} Z'_b X_b(X'_a X_a)^{-1} X'_a e_{1a} \mid x_a, x_b, y_{1a}\}$$

and

$$\begin{aligned} (Z'_b Z_b)^{-1} Z'_b X_b(X'_a X_a)^{-1} X'_a e_{1a} &= (Z'_b Z_b)^{-1} Z'_b \{X_b(X'_a X_a)^{-1} X'_a (y_{1a} - X_a \phi)\} \\ &= (Z'_b Z_b)^{-1} Z'_b X_b(\hat{\phi}_a - \phi). \end{aligned}$$

This term has zero expectation asymptotically because $n_b^{-1} Z'_b Z_b$ and $n_b^{-1} Z'_b X_b$ are bounded in probability and $(\hat{\phi}_a - \phi)$ converges to zero.

B. Variance estimation

Let the parameter of interest be defined by the solution to $U_N(\eta) = \sum_{i=1}^N U(\eta; y_{1i}, y_{2i}) = 0$. We assume that $\partial U_N(\eta)/\partial \theta = 0$. Thus, parameter η is priori independent of θ which is the parameter in the data-generating distribution of (x, y_1, y_2) .

Under the setup of Section 3, let $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)$ be the MLE of $\theta = (\theta_1, \theta_2)$ obtained by solving (7). Also, let $\hat{\eta}$ be the solution to $\bar{U}(\eta \mid \hat{\theta}) = 0$ where

$$\bar{U}(\eta \mid \theta) = \sum_{i \in B} \sum_{j=1}^m w_{ib} w_{ij}^* U(\eta; y_{1i}^{*(j)}, y_{2i}),$$

and

$$w_{ij}^* \propto f(y_{1i}^{*(j)} \mid x_i; \hat{\theta}_1) f(y_{2i} \mid y_{1i}^{*(j)}; \hat{\theta}_2) / h(y_{1i}^{*(j)} \mid x_i)$$

with $\sum_{j=1}^m w_{ij}^* = 1$. Here, $h(y_1 | x)$ is the proposal distribution of generating imputed values of y_1 in the parametric fractional imputation. By introducing the proposal distribution h , we can safely ignore the dependence of imputed values $y_{1i}^{*(j)}$ on the estimated parameter value $\hat{\theta}_1$.

By Taylor linearization,

$$\bar{U}(\eta | \hat{\theta}) \cong \bar{U}(\eta | \theta) + (\partial \bar{U} / \partial \theta'_1) (\hat{\theta}_1 - \theta_1) + (\partial \bar{U} / \partial \theta'_2) (\hat{\theta}_2 - \theta_2)$$

Note that

$$\hat{\theta}_1 - \theta_1 \cong \{I_1(\theta_1)\}^{-1} S_1(\theta_1)$$

where $I_1(\theta_1) = -\partial S_1(\theta_1) / \partial \theta'_1$. Also,

$$\hat{\theta}_2 - \theta_2 \cong \left\{ -\frac{\partial}{\partial \theta'_2} \bar{S}_2(\theta) \right\}^{-1} \bar{S}_2(\theta)$$

where

$$\bar{S}_2(\theta) = \sum_{i \in B} \sum_{j=1}^m w_i w_{ij}^*(\theta) S_2(\theta_2; y_{1i}^{*(j)}, y_{2i}).$$

Thus, we can establish

$$\bar{U}(\eta | \hat{\theta}) \cong \bar{U}(\eta | \theta) + K_1 S_1(\theta_1) + K_2 \bar{S}_2(\theta),$$

where $K_1 = D_{21} I_{11}^{-1}$ and $K_2 = D_{22} I_{22}^{-1}$ with $I_{11} = -E(\partial S_1 / \partial \theta'_1)$, $I_{22} = -E(\partial \bar{S}_2 / \partial \theta'_2)$, $D_{21} = E\{U(\eta) S_1(\theta_1)'\}$ and $D_{22} = E\{U(\eta) S_2(\theta_2)'\}$, we have

$$V\{\bar{U}(\eta | \hat{\theta})\} = \tau^{-1} \{V_1 + V_2\} \tau^{-1'}$$

where $\tau = -E\{\partial \bar{U}(\eta | \theta) / \partial \eta'\}$,

$$V_1 = V \left\{ \sum_{i \in B} w_i (\bar{u}_i^* + K_2 S_{2i}^*) \right\},$$

$\bar{u}_i^* = E[U(\hat{\eta}; y_{1i}, y_{2i}) | y_{2i}; \hat{\theta}]$, and $V_2 = V\{K_1 \sum_{i \in A} w_i S_{1i}\}$. A consistent estimator of each component can be developed similarly to Section 3.

C. Score Tests

In some applications related to measurement error, an analytical question of interest may be phrased in terms of a null hypothesis about the parameter θ . Suppose that $\theta = (\theta_1, \theta_2)$, and the null hypothesis of interest is $H_0 : \theta_2 = \theta_{2,0}$ for a specified $\theta_{2,0}$. Hypotheses about functions of θ_1 and θ_2 can be expressed as a null hypothesis about a sub-vector of interest after appropriate reparametrization. We define a score test using the approach of Rao et al. (1998) and Boos (1992).

Let

$$U_{1i}(\theta_1, \theta_2, \eta) = (U_{11i}(\theta_1, \theta_2, \eta), U_{12i}(\theta_1, \theta_2, \eta)), \quad (\text{A.1})$$

where $U_{1ki}(\theta_1, \theta_2, \eta) = E[S_{1ki}(\theta_1, \theta_2, x_i) | y_i, z_i, \eta]$ for $k = 1, 2$,

$$S_{1i}(\theta) = (S_{11i}(\theta_1, \theta_2, x_i), S_{12i}(\theta_1, \theta_2, x_i)), \quad (\text{A.2})$$

and S_{1ki} is the vector of derivatives of the complete data log likelihood with respect to θ_k . Under the null hypothesis, an estimator $\tilde{\theta}_1$ satisfies,

$$U_{11}(\tilde{\theta}_1, \theta_2, \eta) = \sum_{i \in B} w_{iB} U_{11i}(\tilde{\theta}_1, \theta_{2,0}, \hat{\eta}) = 0, \quad (\text{A.3})$$

and we use parametric fractional imputation to solve (A.3). By a Taylor expansion,

$$\begin{aligned} 0 &= U_{11}(\tilde{\theta}_1, \theta_{2,0}, \hat{\eta}) \\ &\approx U_{11}(\theta_1, \theta_{2,0}, \eta) + \tau_{1,11}(\tilde{\theta}_1 - \theta_1) + \Delta_{1,\eta}(\hat{\eta} - \eta), \end{aligned} \quad (\text{A.4})$$

and

$$U_{12}(\tilde{\theta}_1, \theta_{2,0}, \hat{\eta}) \approx U_{12}(\theta_1, \theta_{2,0}, \eta) + \tau_{1,21}(\tilde{\theta}_1 - \theta_1) + \Delta_{2,\eta}(\hat{\eta} - \eta), \quad (\text{A.5})$$

where $\tau_{1,k1}$ is the matrix of derivatives of $U_{1k}(\theta_1, \theta_{2,0}, \eta)$ with respect to θ_1 , and $\Delta_{k,\eta}$ is the matrix of derivatives of $U_{1k}(\theta_1, \theta_{2,0}, \eta)$ with respect to η . Solving (A.4) for $\tilde{\theta}_1 - \theta_1$ and plugging the resulting expression into (A.5) gives,

$$\begin{aligned} U_{12}(\tilde{\theta}_1, \theta_{2,0}, \hat{\eta}) &= U_{12}(\theta_1, \theta_{2,0}, \eta) - \tau_{1,21} \tau_{1,11}^{-1} \{U_{11}(\theta_1, \theta_{2,0}, \eta)\} \\ &\quad + (-\tau_{1,11}^{-1} \Delta_{1,\eta}, \Delta_{2,\eta})(\hat{\eta} - \eta). \end{aligned} \quad (\text{A.6})$$

An estimate of the variance of $U_{12}(\tilde{\theta}, \theta_{2,0}, \hat{\eta})$ is

$$\hat{V}_s = \hat{V} \left\{ \sum_{i \in B} w_{iB} \hat{v}_i \right\} + (-\hat{\tau}_{1,11}^{-1} \hat{\Delta}_{1,\eta}, \hat{\Delta}_{2,\eta}) \hat{V} \{ \hat{\eta} \} (-\hat{\tau}_{1,11}^{-1} \hat{\Delta}_{1,\eta}, \hat{\Delta}_{2,\eta})', \quad (\text{A.7})$$

where

$$\hat{v}_i = U_{12i}(\tilde{\theta}, \theta_{2,0}, \hat{\eta}) - \tau_{1,21} \tau_{1,11}^{-1} U_{11i}(\tilde{\theta}_1, \theta_{2,0}, \eta). \quad (\text{A.8})$$

A size α score test of the null hypothesis, $H_0 : \theta_2 = \theta_{2,0}$ rejects if $T(\theta_{2,0}) > \chi_p^2(1 - \alpha)$, where

$$T(\theta_{2,0}) = [U_{12}(\tilde{\theta}_1, \theta_{2,0}, \hat{\eta})]' \hat{V}_s^{-1} [U_{12}(\tilde{\theta}_1, \theta_{2,0}, \hat{\eta})], \quad (\text{A.9})$$

and $\chi_p^2(\cdot)$ is the quantile function of a chi-squared distribution with p degrees of freedom. A confidence region for θ_2 with confidence level $1 - \alpha$ is the set of θ_2 with $T(\theta_{2,0} = \theta_2) < \chi_p^2(1 - \alpha)$.

References

- BAKER, K. H., HARRIS, P., & O'BRIEN, J. (1989). Data fusion: An appraisal and experimental evaluation. *Journal of the Market Research Society* **31** 152–212.
- BEAUMONT, J. F., BOCCI, C. (2009). Variance estimation when donor imputation is used to fill in missing values. *Can. J. Statist.* **37** 400–416.
- BOOS, D. D. (1992). On generalized score tests. *Am. Stat.* **46** 327–333.
- CHEN, J. & SHAO, J. (2001). Jackknife variance estimation for nearest neighbor imputation. *J. Am. Statist. Assoc.* **96** 260–269.
- CHIB, S. & GREENBERG, E. (1995). Jackknife variance estimation for nearest neighbor imputation. *Am. Stat.* **46** 327–333.
- CHIPPERFIELD, J. O. & STEEL, D. G. (2009). Design and estimation for split questionnaire surveys. *J. Offic. Statist.* **25** 227–244.

- D’ORAZIO, M., ZIO, M. D. & SCANU, M. (2006). *Statistical Matching: Theory and Practice*. Chichester, UK: Wiley.
- FULLER, W. A. (2009). *Sampling Statistics*, Hoboken, NJ: John Wiley & Sons, Inc.
- GELMAN, A., CARLIN, J.B., STERN, H.S., RUBIN, D.B. (2003). *Bayesian Data Analysis*, Chapman and Hall Texts in Statistical Science. Chapman and Hall/CRC, second edition.
- GONZALEZ, J. & ELTINGE, J. (2008). Adaptive matrix sampling for the consumer expenditure quarterly interview survey. In *Proc. Survey Res. Meth. Sect.* Washington DC: American Statistical Association, 2081–2088.
- GUO, Y. & LITTLE, R. J. (2011). Regression analysis with covariates that have heteroskedastic measurement error. *Statist. Med.* **30** 2278–2294.
- HAZIZA, D. (2009). Imputation and inference in the presence of missing data. In *Handbook of Statistics, Volume 29, Sample Surveys: Theory Methods and Inference*, Editors: C.R. Rao and D. Pfeffermann, 215-246.
- HERZOG, T. N & SCHEUREN, F. J & WINKLER, W.E. *Data Quality and Record Linkage Techniques*. New York: Springer
- JBRAHIM, J. G. *Incomplete data in generalized linear models Journal of the American Statistical Association* **85** 765-769.
- KIM, J. K. (2011). Parametric fractional imputation for missing data analysis. *Biometrika* **98** 119–132.
- KIM, J. K. & RAO, J. N. K. (2012). Combining data from two independent surveys: a model-assisted approach. *Biometrika* **99** 85–100.
- KIM, J. K. & YANG, S. (2013). Fractional hot deck imputation for robust inference in survey sampling. *Survey Methodology*, Accepted for publication.

- KIM, J. K. & SHAO, J. (2013). *Statistical Methods in Handling Incomplete Data*, Chapman and Hall / CRC.
- LAHIRI, P. & LARSEN, M.D. (2005). Regression analysis with linked data. *Journal of the American Statistical Association* **100** 1265–1275.
- LEULESCU, A. & AGAFITEI, M. (2013). Statistical matching: a model based approach for data integration. *Eurostat Methodologies and Working Papers*
- MORGAN, S. L. & WINSHIP, C. (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. New York, USA: Cambridge University Press.
- MORIARITY, C. & SCHEUREN, F. (2001). Statistical matching: a paradigm for assessing the uncertainty in the procedure. *J. Offic. Statist.* **17** 407–422.
- PLUMMER, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling, In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*.
- RAGHUNATHAN, T. E. & GRIZZLE, J. E. (1995). A split questionnaire design. *J. Am. Statist. Assoc.* **90** 54–63.
- RAO, J. N. K., SCOTT, A. J., & SKINNER, J. C. (1998). Quasi-score tests with survey data. *Statist. Sinica* **8** 1059–1070.
- RÄSSLER, S. (2002). *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*. New York, NY: Springer-Verlag.
- RIDDER, S. & MOFFIT, R. (2007). The econometrics of data combination. *Handbook of Econometrics* 5470–5544.